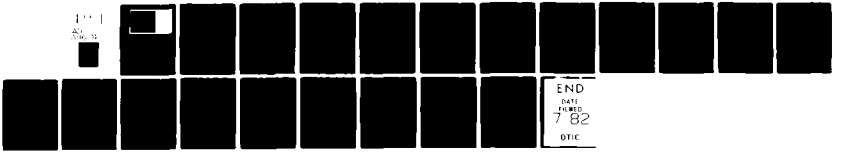
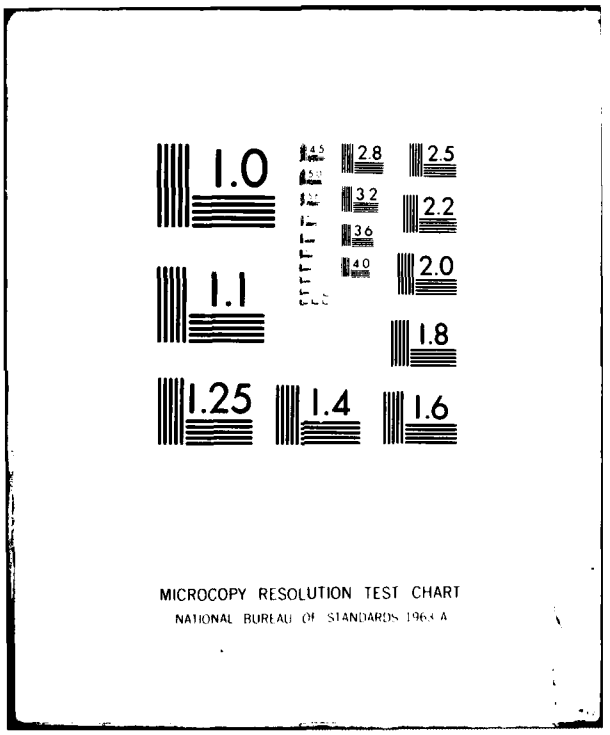
AD-A116 174 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
KERNEL-BASED DENSITY ESTIMATION USING CENSORED, TRUNCATED OR GR--ETC(U)
MAY 82 D M TITTERINGTON DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2382 NL



END
DATE FILMED
7 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

MRC Technical Summary Report # 2382

KERNEL-BASED DENSITY ESTIMATION USING CENSORED, TRUNCATED OR GROUPED DATA

D. M. Titterington

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

May 1982

(Received March 16, 1982)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

KERNEL-BASED DENSITY ESTIMATION USING CENSORED, TRUNCATED OR GROUPED DATA

D. M. Titterington*

Technical Summary Report #2382

May 1982

ABSTRACT

Censoring, truncation and grouping represent different but related forms of incompleteness. Methods of producing kernel functions on the incomplete observations are proposed. They involve substituting for or averaging over the incomplete observations. Consistency of the procedures in terms of the criterion of integrated mean squared error is established and optimal choice of smoothing parameter is achieved.

AMS (MOS) Subject Classifications: Primary: 62G05, Secondary: 62-07.

Key Words: density estimation, kernel method, censoring, truncation, grouping, consistency, imputation.

Work Unit Number 4 - Statistics and Probability

* Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

SIGNIFICANCE AND EXPLANATION

When data are used to estimate a probability density function, either a special parametric form is assumed for the latter, a Normal density being a common particular case, or a nonparametric method is employed. One such example is the Kernel method.

For many problems data are available which are incomplete in some sense. Three types of incompleteness are censoring (in which the exact values of some observations are unknown) truncation (in which the data are known exactly and also to be restricted to a certain range) and grouping, of which one manifestation is data in the form of a histogram.

The basic kernel method relies on the data being "complete" and this paper gives adaptations to cope with the above three types of incompleteness. One feature of density estimation by the kernel method is the need to choose, in some sensible or, if possible, optimal way, a parameter which dictates the smoothness of the resulting estimate. A formula is derived for the value of the smoothing parameter which is optimal according to one particular criterion.

Techniques for coping with incomplete data within parametric models are well established. It is important to deal with such problems with nonparametric methods as well because, although no parametric model may be correct for a given application, the converse is true for nonparametric methods, at least asymptotically.

KERNEL-BASED DENSITY ESTIMATION USING CENSORED, TRUNCATED OR GROUPED DATA

D. M. Titterington*

1. INTRODUCTION

The problem of density estimation using censored, truncated or grouped data is an important one. When a parametric model is acceptable, the problem becomes one of parameter estimation and the maximum likelihood approach is dealt with succinctly by Dempster et al (1977, Section 4.2). A maximum likelihood approach to the nonparametric version of the problem is dealt with by Turnbull (1974, 1976). This does not, however, lead to a smooth estimate of the underlying probability density function. The object of this paper is to propose methods for achieving this aim based on the kernel approach and to investigate some asymptotic properties.

One condition has to be imposed, however, namely that some information about the overall density be available. In the parametric case this is supplied by the parametric family chosen. In the absence of this we shall require that a set of n_0 observations be available which are quite unaffected by the censoring, truncation or grouping mechanism. (If, for instance, only grouped data are available, in the form of a histogram with fixed bin size, then there is no hope of consistently estimating the density everywhere without further information.) The incomplete data, therefore, may be regarded as supplementary to the original n_0 observations.

* Department of Statistics, University of Glasgow, Glasgow G12 8QW, Scotland.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

The methodology will be similar to that of Titterington and Mill (1981) who dealt with multivariate data with missing values. In what follows, only univariate continuous data are considered.

2. THE DATA AVAILABLE

2.1 Censoring

Along with n_0 independent observations $\underline{x} = (x_1,\ldots,x_{n_0})$ from the underlying distribution on a sample space, X, with probability density function $f(\cdot)$, we have $\underline{y} = (y_1,\ldots,y_{n_1})$, n_1 independent observations, with known values, in A, a subset of X, and n_2 independent observations known to be in $\bar{A}$, the complement of A in X.

We assume that, given $n_1 + n_2$, $n_1 \sim Bi(n_1 + n_2, P(A))$, where

$$P(A) = \int_A f(x)dx \quad ,$$

and that $n_0 = \theta_0(n_0 + n_1 + n_2)$.

(The asymptotic results we obtain would hold also under the assumption that given $n_0 + n_1 + n_2 = n$, $n_0 \sim Bi(n, \theta_0)$.)

2.2 Truncation

Along with $\underline{x}$ we have $\underline{y} = (y_1,\ldots,y_{n_1})$, n_1 independent observations from A, a subset of X. The p.d.f. for each of the $\{y_j\}$ is therefore

$$f(y)/P(A) \qquad (y \in A) \quad .$$

2.3 Grouping

Along with $\underline{x}$ we have independent samples of sizes $n_1,\ldots,n_m$, containing independent observations from members $A_1,\ldots,A_m$, respectively, of a partition of X. Given $n_1 +\ldots+ n_m$, the n_j's are multinomial, with cell probabilities $\{P(A_j)\}$, where

$$P(A_j) = \int_{A_j} f(x)dx, \quad j = 1,\ldots,m \quad .$$

3. KERNEL-BASED DENSITY ESTIMATION WITH INCOMPLETE DATA

Given a data-set $\underline{t} = (t_1,\ldots,t_n)$ of n independent identically distributed observations, each with p.d.f. $f(x)$, $x \in X$, a kernel-based density estimate of $f(x)$ takes the form

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K((x-t_i)/h) \quad ,$$

where h is a smoothing parameter and the kernel function, $K(\cdot)$, is itself a density, usually with its mode at zero. We shall assume that $K(\cdot)$ is square-integrable and symmetric, with bounded first and second absolute moments. Define

$$I_1 = \int u^2 K(u)du$$

and

$$I_2 = \int K^2(u)du \quad .$$

One interpretation of our basic question is to ask what to use for the kernel function on an incomplete observation. In the spirit of Titterington and Mill (1981) we propose two possible solutions.

(A) Plug in a "complete" data point for the incomplete one.

(B) Average out the "incompleteness".

In the case of censored data, for instance, we require kernel functions on the n_2 censored observations in $\bar{A}$. The corresponding p.d.f. is

$$f(z)/P(\bar{A}) \qquad (z \in \bar{A}) \quad ,$$

where $P(\bar{A}) = 1 - P(A)$.

Although this density is unknown we do have, from $\underline{x}$, an estimate

$$\hat{f}_0(z)/\hat{P}(\bar{A}), \qquad (z \in \bar{A}) \quad , \tag{1}$$

where

$$\hat{f}_0(z) = (n_0 h)^{-1} \sum_{i=1}^{n_0} K((z-x_i)/h)$$

and

$$\hat{P}(\bar{A}) = \int_{\bar{A}} \hat{f}_0(z)dz \quad .$$

As justified in Titterington and Mill (1981) we dismiss, for (A), the "deterministic" mean-imputation procedure of plugging in the expected value from (1) for each censored observation. Instead we use simulated values from (1). We may use one value or, more generally, r independent values, $z_{i1},\ldots,z_{ir}$, for the ith censored observation, giving the "kernel"

$$(rh)^{-1} \sum_{j=1}^{r} K((x-z_{ij})/h) \quad .$$

It is natural that the averaging in method (B) be carried out using (1), giving the following "kernel" on each censored observation.

$$\{h\hat{P}(\bar{A})\}^{-1} \int_{\bar{A}} K((x-z)/h)\hat{f}_0(z)dz \quad . \qquad (2)$$

In practice this integral may well have to be evaluated numerically, in contrast to what is possible in the missing-values problem (Titterington and Mill, 1981). Direct simulation from (1) will also be awkward but here the problem is eased in practice if we simulate from the density $\hat{f}_0(z)$ $(z \in X)$, which is a mixture density that should be easy for simulation, and ignore all values not in $\bar{A}$.

For truncated data, the "incomplete" observations are not so immediately apparent. We introduce them deviously, as in Dempster et al (1977), by proposing that, corresponding to the n_1 truncated observations, $\underline{y}$, there lurk n_2 observations in $\bar{A}$ to make up $n_1 + n_2$ altogether in X. Given n_1, n_2 has a negative binomial distribution on $(0,1,2,\ldots)$, with

$$E(n_2|n_1) = n_1 P(\bar{A}).P(A)^{-1} \quad .$$

For each of these n_2 we generate kernels, as above, by simulating or averaging. Joint simulation of n_2 and the corresponding $\{z_{ij}\}$ is neatly achieved by simulating from the p.d.f. $\hat{f}_0(z)$ $(z \in X)$ until rn_1 values in

A have been generated and by regarding the remainder as the rn_2 extra values in $\bar{A}$.

It should now be clear how to deal with grouped data, so that we may list the following proposals for density estimates.

3.1 Censoring

$$\text{(A)} \qquad \hat{f}_A(x) = (n_0+n_1+n_2)^{-1}h^{-1}\left\{\sum_{i=1}^{n_0} K((x-x_i)/h) + \sum_{i=1}^{n_1} K((x-y_i)/h) + r^{-1}\sum_{i=1}^{rn_2} K((x-z_i)/h)\right\}, \tag{3}$$

where $(z_1,\ldots,z_{rn_2})$ denote the simulated values, a notation which fits in better with the truncation case.

$$\text{(B)} \qquad \hat{f}_B(x) = (n_0+n_1+n_2)^{-1}h^{-1}\left\{\sum_{i=1}^{n_0} K((x-x_i)/h) + \sum_{i=1}^{n_1} K((x-y_i)/h) + n_2\hat{P}(\bar{A})^{-1}\int_{\bar{A}} K((x-z)/h)\hat{f}_0(z)dz\right\}. \tag{4}$$

3.2 Truncation

Formulae (3) and (4) are again relevant. It must be remembered that, given n_1, n_2 is the realization of a negative binomial random variable, as discussed above.

3.3 Grouped data

$$\text{(A)} \qquad \hat{f}_A(x) = \Big(n_0 + \sum_{k=1}^{m} n_k\Big)^{-1}h^{-1}\left\{\sum_{i=1}^{n_0} K((x-x_i)/h) + r^{-1}\sum_{k=1}^{m}\sum_{i=1}^{n_k}\sum_{j=1}^{r} K((x-z_{ij}^{(k)})/h)\right\}. \tag{5}$$

$$\text{(B)} \qquad \hat{f}_B(x) = \Big(n_0 + \sum_{k=1}^{m} n_k\Big)^{-1}h^{-1}\left\{\sum_{i=1}^{n_0} K((x-x_i)/h) + \sum_{k=1}^{m} n_k\hat{P}(A_k)^{-1}\int_{A_k} K((x-z)/h)\hat{f}_0(z)dz\right\}. \tag{6}$$

In (5), the $\{z_{ij}^{(k)}\}$ are independent, with p.d.f.'s

$$\hat{f}_0(z)/\hat{P}(A_k) \quad (z \in A_k), \quad \text{for each } i,j,k \ ,$$

with

$$\hat{P}(A_k) = \int_{A_k} \hat{f}_0(z)dz \ .$$

4. ASYMPTOTIC RESULTS

In this section we establish consistency of the density estimators under the criterion of integrated mean squared error and derive optimal values for the smoothing parameter, h. Specifically, we show that, for suitably defined n,

$$\int \mathbb{E}\{\hat{f}(x)-f(x)\}^2dx = \frac{1}{4} H^2h^4 + Gn^{-1}h^{-1} + o(h^4+n^{-1}h^{-1}) \ , \qquad (7)$$

for certain constants H and G. The dominant terms in (7) are minimized by

$$h^* = (GH^{-2}n^{-1})^{1/5} \qquad (8)$$

and, under this choice, the right hand side of (7), of order $O(n^{-4/5})$, tends to zero as $n \to \infty$. The calculations involved are similar to those of Rosenblatt (1956) and Epanechnikov (1969).

Note that

$$\int \mathbb{E}(f(x) - \hat{f}(x))^2dx = \int (\mathbb{E}\hat{f}(x) - f(x))^2dx + \int \text{var } \hat{f}(x)dx \ . \qquad (9)$$

The dominant terms in (7) come from these two constituent parts, which we evaluate below. For all three types of incompleteness we may observe that, conditioning on $\underline{x}$, $\underline{y}$ (in the cases of censoring and truncation) and all sample sizes $\underline{n}$, averaging over the simulated data, $\underline{z}$, gives

$$\mathbb{E}_{\underline{z}} \hat{f}_A(x) = \hat{f}_B(x) \ , \quad \text{for all } x \ .$$

Thus, unconditionally,

$$\mathbb{E}\hat{f}_A(x) = \mathbb{E}\hat{f}_B(x)$$

and

$$\text{var } \hat{f}_A(x) = \mathop{\mathbb{E}}_{\underline{x},\underline{y},\underline{n}} \text{var } \hat{f}_A(x) + \mathop{\text{var}}_{\underline{x},\underline{y},\underline{n}} \hat{f}_B(x) \ . \qquad (10)$$

Thus, almost certainly, $\hat{f}_A(\cdot)$ will not be as efficient as $\hat{f}_B(\cdot)$, although, as we shall see, its comparative ease of application may make it the preferred method in practice.

In the Appendix, the case of censored data is dealt with in detail, with the results that H^2 is the same for both method (A) and method (B), whereas the values of G are different. G_A and G_B are given by equations (A.8) and (A.7).

Exactly the same results will hold for the truncated-data case except that the value n in (7) has to be interpreted differently. In practice $N = n_0 + n_1$ will be known and n is to be interpreted as the total sample size, given n_1 and n_0. Thus, formulae in terms of N can be obtained by substituting from

$$n = N\{\theta_0 + (1-\theta_0)/P(A)\}$$

in the results for censored data.

Calculations for the case of grouped data give

$$E\hat{f}_A(x) = E\hat{f}_B(x) = f(x) + b_h(x)$$

$$+ (1-\theta_0)\{b_h(x) - P(A(x))^{-1}B_h(A(x))f(x)\} + o(h^2) \quad ,$$

where $A(x)$ is the grouping interval containing x.

In the third term we have followed the approximation leading to equations (A.4). Also,

$$\int \text{var}\, \hat{f}_B(x)dx = (nh)^{-1}\{\theta_0 I_2 + 2(1-\theta_0)I_4 + (1-\theta_0)^2 I_3/\theta_0\} + o(n^{-1}h^{-1}) \quad ,$$

where I_4 and I_3 are defined in the Appendix.

$$\int \text{var}\, \hat{f}_A(x)dx = \int \text{var}\, \hat{f}_B(x)dx + (nh)^{-1}(1-\theta_0)I_2/r + o(n^{-1}h^{-1}) \quad .$$

5. SOME NUMERICAL RESULTS

When there is a substantial amount of censored or truncated data to supplement $\underline{x}$, the density estimator which incorporates them should be better than that based on $\underline{x}$. We present some numerical results for the case of censored data from a standard Normal distribution, using a Normal kernel function, for which $I_1 = 1$, $I_2 = (2\sqrt{\pi})^{-1}$, $I_3 = (\sqrt{6\pi})^{-1}$ and $I_4 = \frac{1}{2} I_2$.

With the optimal choice, h^*, for the smoothing parameter, the dominant term in (7) is

$$S \propto (G^2 H n^{-2})^{2/5} \quad .$$

If only the complete data are used, then the corresponding value is

$$S_0 \propto (G_0^2 H_0 n_0^{-2})^{2/5} \quad ,$$

where, effectively, $n_0 = n\theta_0$, $G_0 = I_2$ and

$$H_0^2 = I_1^2 \int_{-\infty}^{\infty} \{f''(x)\}^2 dx \quad .$$

Of interest is the ratio

$$R = (S/S_0)^{5/2} = G^2 H \theta_0^2 / G_0^2 H_0 \quad .$$

Since $G_0 = I_2$,

$$R = F^2 H \theta_0^2 / H_0 \quad ,$$

where

$$F_B = \theta_0 + (1-\theta_0)\{P(A) + P(\bar{A})(\theta_0^{-1}(1-\theta_0)I_3/I_2 + 2I_4/I_2)\}$$

and $F_A = F_B + (1-\theta_0)P(\bar{A})/r$.

As an illustrative simple example take $A = (-\infty, 0)$. Then, from the Appendix,

$$H^2 = \int_{-\infty}^{\infty} H^2(x) dx \quad ,$$

where

$$H(x) = I_1 f''(x) \qquad (x \in A)$$

$$= I_1\{(2-\theta_0)f''(x) - \frac{1}{2} f(x) \int_0^{\infty} f''(y)dy\} \qquad (x \in \bar{A}) \quad .$$

Since $\int_0^\infty f''(x) = 0$, we obtain

$$H^2 = H_0^2\{1 + (2-\theta_0)^2\}/2 \quad ,$$

so that

$$R = F^2\theta_0^2\sqrt{\{1 + (2-\theta_0)^2\}/2} \quad .$$

In particular, since $P(A) = \frac{1}{2}$,

$$F_B = \theta_0 + (1-\theta_0)\{1 + \theta_0^{-1}(1-\theta_0)/\sqrt{6}\} = 1 + \theta_0^{-1}(1-\theta_0)^2/\sqrt{6}$$

$$F_A = F_B + (1-\theta_0)/2r \quad .$$

Thus

$$R_B = \{\theta_0 + (1-\theta_0)^2/\sqrt{6}\}^2[\{1 + (2-\theta_0)^2\}/2]^{1/2}$$

$$R_A = \{\theta_0 + (1-\theta_0)^2/\sqrt{6} + \theta_0(1-\theta_0)/r\}^2[\{1 + (2-\theta_0)^2\}/2]^{1/2} \quad .$$

Values of R_A for various values of r and θ_0 are displayed in Table 1. The row for $r = \infty$ corresponds to R_B. R_A can be close to R_B for only a small value of r, a phenomenon reported also by Titterington and Mill (1981). Thus, although method (B) is in principle to be preferred, method (A) can easily be almost as good, as well as being much easier to apply.

6. DISCUSSION

We end with the following comments.

(i) Although the censoring and truncation requirements are very simple, there is difficulty in extending the analysis to more complicated ones, based on a partition of X, on the lines of Dempster et al (1977, Section 4.2).

(ii) In practice some data-based method may be required for choosing the smoothing parameter, h. The formula given by (8) depends on the unknown density itself. A useful reference is Scott and Factor (1981).

(iii) It has to be admitted that some of the gains embodied in Table 1 are not remarkable. However, the methods of the paper should be valuable as nonparametric imputation procedures (particularly method (A) with $r = 1$). In many sample survey projects with non-response it is desirable to impute the missing values in such a way as to provide a "fair" complete data-set. Given that the statistical characteristics underlying the incompleteness process are as described in the paper, method (A) will certainly achieve this aim.

(iv) Only the case of <u>fixed</u> Type I censoring has been considered here. The same methods can be applied to random censoring and consistency will obtain, provided we have a data-set D_0 which is known not to have been subject to the possibility of censoring. In many problems involving random censoring such a D_0 is not available and to use the uncensored data we have, which would correspond to D_1, for imputation or averaging would almost certainly lead to bias. For this case methods have been developed for smoothing the nonparametric Kaplan-Meier estimate of the survival curve; see Foldes and Retjo (1981) and Yandell (1981).

TABLE 1

Some values of R_A for the Example in Section 5

r	θ_0 0.1	0.3	0.5	0.7	0.9
1	0.41	0.70	0.93	1.04	1.04
2	0.34	0.51	0.67	0.82	0.95
5	0.31	0.41	0.54	0.70	0.89
10	0.29	0.39	0.50	0.67	0.88
∞ (R_B)	0.28	0.35	0.46	0.63	0.86

APPENDIX. CALCULATION OF INTEGRATED MEAN SQUARED ERROR FOR CENSORED DATA CASE.

Once the first term on the right hand side of (10) is dealt with, the remaining calculations are all related to $\hat{f}_B(x)$ as given by (4).

From (3),

$$\underset{\underline{z}}{\text{var}}\, \hat{f}_A(x) = (n_0+n_1+n_2)^{-2} n_2 \{r\hat{P}(\bar{A})\}^{-1} \{h^{-2} \int_{\bar{A}} K^2((x-z)/h)\hat{f}_0(z)dz + o(h^{-2})\}$$

$$= (n_0+n_1+n_2)^{-2} n_2 \{rP(\bar{A})\}^{-1} \{h^{-2} \int_{\bar{A}} K^2((x-z)/h)f(z)dz + o(h^{-2})\} \quad .$$

If, given $n_1 + n_2$, $n_2 \sim Bi(n_1+n_2, P(\bar{A}))$ and if, given $n_0+n_1+n_2 = n$, $n_0 = \theta_0 n$ (or $n_0 \sim Bi(n, \theta_0)$), then the dominant term in $\mathbf{E}\, \underset{\underline{z}}{\text{var}}\, \hat{f}_A(x)$, for use in (10), is

$$(1-\theta_0)(nrh^2)^{-1} \int_{\bar{A}} K^2((x-z)/h)f(z)dz \quad . \tag{A.1}$$

(An unqualified "E" or "var" will be assumed to involve averaging over any random variation not so far acounted for.)

This leads to the following contribution to (9).

$$\int \mathbf{E}\, \underset{\underline{z}}{\text{var}}\, \hat{f}_A(x)dx \approx (1-\theta_0)(nrh^2)^{-1} \int_X \int_{\bar{A}} K^2((x-z)/h)f(z)dzdx \quad .$$

Substitution of x by $z - uh$ gives

$$(1-\theta_0)(nrh)^{-1} \int_{\bar{A}} f(z)dz \int K^2(u)du$$

$$= (1-\theta_0)(nrh)^{-1} P(\bar{A}) I_2 \quad . \tag{A.2}$$

We now concentrate on $\hat{f}_B(x)$ from (4).

In the notation of Silverman (1978),

$$\hat{f}_0(x) = f(x) + b_h(x) + \sigma_h(x) \quad ,$$

where

$$b_h(x) = \frac{1}{2} h^2 I_1 f''(x) + o(h^2)$$

and $\sigma_h(x)$ is a zero-mean Gaussian process with variance function $(n_0h)^{-1}I_2f(x) + o(n_0^{-1}h^{-1})$, given n_0. Then

$$\hat{P}(\bar{A}) = P(\bar{A}) + B_h(\bar{A}) + S_h(\bar{A}) \quad ,$$

where

$$B_h(\bar{A}) = \int_{\bar{A}} b_h(x)dx$$

and

$$S_h(\bar{A}) = \int_{\bar{A}} \sigma_h(x)dx \quad .$$

Thus

$$\{h\hat{P}(\bar{A})\}^{-1} \int_{\bar{A}} K((x-z)/h)\hat{f}_0(z)dz$$

$$\approx \{hP(\bar{A})\}^{-1}\{1 - (B_h(\bar{A})+S_h(\bar{A}))/P(\bar{A})\} \int_{\bar{A}} K((x-z)/h)(f(z)+b_h(z)+\sigma_h(z))dz$$

$$\approx P(\bar{A})^{-1}\{h^{-1} \int_{\bar{A}} K((x-z)/h)f(z)dz + \beta_h(x) + \gamma_h(x)\} \quad ,$$

where

$$\beta_h(x) = h^{-1} \int_{\bar{A}} K((x-z)/h)b_h(z)dz - B_h(\bar{A})\{hP(\bar{A})\}^{-1} \int_{\bar{A}} K((x-z)/h)f(z)dz \quad (A.3)$$

and

$$\gamma_h(x) = h^{-1} \int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz - S_h(\bar{A})\{hP(\bar{A})\}^{-1} \int_{\bar{A}} K((x-z)/h)f(z)dz \quad .$$

When taking expectations over the sample sizes, the dominant term is obtained simply by inserting expected values, $n\theta_0$ for n_0, $n(1-\theta_0)P(A)$ for n_1 and $n(1-\theta_0)P(\bar{A})$ for n_2. Variances over the sample sizes will be of order $O(n^{-1})$, which is $o(n^{-1}h^{-1})$ and $o(h^2)$, for h of the order we shall use, namely $O(n^{-1/5})$. These variances may therefore be neglected.

It follows that, if all but the dominant terms are neglected,

$$E\,\hat{f}_B(z) = n^{-1}[n\theta_0\{f(x)+b_h(x)\} + n(1-\theta_0)P(A)\{hP(A)\}^{-1}$$

$$\int_A K((x-y)/h)f(y)dy + n(1-\theta_0)P(\bar{A})\{P(\bar{A})\}^{-1}\{h^{-1} \int_{\bar{A}} K((x-y)/h)f(y)dy$$

$$+ \beta_h(x)\}]$$

$$= n^{-1}[n\theta_0(f(x)+b_h(x)) + n(1-\theta_0)h^{-1}\int K((x-y)/h)f(y)dy$$

$$+ n(1-\theta_0)\beta_h(x)]$$

$$= f(x) + b_h(x) + (1-\theta_0)\beta_h(x) + o(h^2) \quad .$$

Note, from (A.3), that $\int \beta_h(x)dx = 0$. Also, for small h, it is approximately true that

$$\beta_h(x) = \begin{cases} 0 & (x \in A) \\ b_h(x) - B_h(\bar{A})P(\bar{A})^{-1}f(x) & (x \in \bar{A}) \end{cases} \quad . \tag{A.4}$$

If we define $H(x)$ by

$$\frac{1}{2}h^2H(x) = b_h(x) + (1-\theta_0)\beta_h(x) \quad , \tag{A.5}$$

then

$$\int \{E\hat{f}_B(x) - f(x)\}^2dx = \frac{1}{4}h^4H^2 + o(h^4) \quad ,$$

where $H^2 = \int H^2(x)dx$. The approximation in (A.4) will be useful in calculating H^2.

The main remaining calculation is to evaluate

$$\underset{\underline{x},\underline{y}}{\text{var}}\ \hat{f}_B(x) \quad ,$$

into which we shall substitute mean values for n_0, n_1 and n_2.

The dominant term in the variance over $\underline{y}$ (which is independent of $\underline{x}$) becomes

$$n^{-1}(1-\theta_0)P(A)h^{-2}\int_A K^2((x-y)/h)f(y)dy \cdot P(A)^{-1}$$

and the integral of this over x is

$$(1-\theta_0)P(A)I_2\, n^{-1}h^{-1} \quad . \tag{A.6}$$

The remaining contribution to (10) is the variance, over $\underline{x}$, of

$$n^{-1}\{n\theta_0\sigma_h(x) + n(1-\theta_0)\gamma_h(x)\}, \text{ that is,}$$

$$\theta_0\sigma_h(x) + (1-\theta_0)h^{-1}\int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz - (1-\theta_0)S_h(\bar{A})\{hP(\bar{A})\}^{-1} \times$$

$$\int_{\bar{A}} K((x-z)/h)f(z)dz \quad .$$

Given n_0, we have the following.

(i) $\text{var}\ \sigma_h(x) = (n_0h)^{-1}I_2f(x) + o(n_0^{-1}h^{-1})$,

so $\int \text{var}\ \sigma_h(x)dx = (n_0h)^{-1}I_2 + o(n_0^{-1}h^{-1})$.

(ii) $\text{var}\{h^{-1}\int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz\}$

$$= \underset{\underline{x}}{E}\{h^{-2}\int_{\bar{A}}\int_{\bar{A}} K((x-z)/h)K((x-y)/h)\sigma_h(z)\sigma_h(y)dydz\}$$

$$\approx n_0^{-1}h^{-4}\{\int f(u)\int_{\bar{A}}\int_{\bar{A}} K((x-z)/h)K((x-y)/h)K((z-u)/h)K(y-u)/h)dydzdu\} \quad .$$

Integrate over x and substitute

$$\frac{x-z}{h} = w;\ \frac{y-u}{h} = v;\ \frac{z-u}{h} = t, \text{ so that } \frac{x-y}{h} = w + t - v \quad .$$

Thus

$$\int \text{var}\{h^{-1}\int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz\}dx$$

$$\approx n_0^{-1}h^{-1}\int_{\bar{A}} f(y)dy \iiint K(v)K(w)K(t)K(w+t-v)dvdtdw$$

$$\approx (n\theta_0h)^{-1}P(\bar{A})I_3 \quad ,$$

where I_3 is the triple integral.

(iii) $\int \text{cov}\{\sigma_h(x), h^{-1}\int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz\}dx$

$$\approx n_0^{-1}h^{-3}\iiint_{\bar{A}} K((x-z)/h)K((x-u)/h)K((z-u)/h)f(u)dzdxdu \quad .$$

Substitute $u = z-vh$, $x = z+wh$. Thus, the right hand side is

$$n_0^{-1}h^{-1}\iint K(w)K(v+w)K(v)dv \cdot \int_{\bar{A}} f(z-vh)dz$$

$$\approx n_0^{-1}h^{-1}I_4\ P(\bar{A}) \approx (nh\theta_0)^{-1}P(\bar{A}) \cdot I_4 \quad ,$$

where I_4 is the double integral.

(iv) $\operatorname{cov}\{\sigma_h(x), S_h(\bar{A})\} = \int_{\bar{A}} \operatorname{cov}(\sigma_h(x), \sigma_h(y))dy$

$$= \int_{\bar{A}} \{n_0^{-1}h^{-2} \int K((x-z)/h)K((y-z)/h)f(z)dz + o(n_0^{-1}h^{-2})\}dy \quad .$$

Substitute $(x-z)/h = u$, $(y-x)/h = v$. Then

$$\operatorname{cov}(\sigma_h(x), S_h(\bar{A})) = O(n_0^{-1}) = O(n^{-1}) = o(n^{-1}h^{-1}) \quad .$$

(v) $\operatorname{var}(S_h(\bar{A})) = \int_{\bar{A}} \int_{\bar{A}} \operatorname{cov}(\sigma_h(y), \sigma_h(z))dydz$

$$\approx n_0^{-1}h^{-2} \int_{\bar{A}} \int_{\bar{A}} \{\int K((y-x)/h)K((z-x)/h)f(x)dx\}dydz$$

$$= O(n_0^{-1}) = o(n^{-1}h^{-1}), \text{ as in (iv).}$$

(vi) $\operatorname{cov}(h^{-1} \int_{\bar{A}} K((x-z)/h)\sigma_h(z)dz, S_h(\bar{A}))$

$$= h^{-1} \int_{\bar{A}} \int_{\bar{A}} K((x-z)/h) \operatorname{cov}(\sigma_h(z), \sigma_h(y))dzdy$$

$$\approx n_0^{-1}h^{-3} \int \{\int_{\bar{A}} \int_{\bar{A}} K((x-z)/h)K((z-u)/h)K(y-u)/h)f(u)dzdy\}du$$

$$= O(n_0^{-1}) = o(n^{-1}h^{-1}), \text{ also.}$$

Thus the dominant term in the integrated variance over x is obtained from (i), (ii) and (iii). We obtain

$$\int \underset{\underline{n}}{E} \operatorname{var}_{\underline{x}} \hat{f}_B(x)dx = (nh)^{-1}\{\theta_0 I_2 + (\frac{(1-\theta_0)^2}{\theta_0} I_3 + 2(1-\theta_0)I_4)P(\bar{A})\} \quad .$$

Combining this with (A.6) we obtain

$$\int \operatorname{var} \hat{f}_B(x)dx = G_B n^{-1}h^{-1} + o(n^{-1}h^{-1}) \quad ,$$

where

$$G_B = I_2\{\theta_0 + (1-\theta_0)P(A)\} + (1-\theta_0)P(\bar{A})\{(1-\theta_0)I_3/\theta_0 + 2I_4\} \quad . \qquad (A.7)$$

With the addition of (A.2), we have

$$\int \operatorname{var} \hat{f}_A(x)dx = G_A n^{-1}h^{-1} + o(n^{-1}h^{-1}) \quad ,$$

where

$$G_A = G_B + (1-\theta_0)P(\bar{A})I_2/r \quad . \qquad (A.8)$$

REFERENCES

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. B, 39, 1-38.

EPANECHNIKOV, V. (1969). Nonparametric estimation of a multidimensional probability density. Theory of Prob. and Applics., 14, 153-158.

FOLDES, A. and REJTO, L. (1981). Strong uniform consistency for nonparametric survival curve estimators from randomly censored data. Ann. Statist., 9, 122-129.

ROSENBLATT, M. (1956). Remarks on some non-parametric estimates of a density function. Ann. Math. Statist., 27, 832-837.

SCOTT, D. W. and FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. J. Amer. Statist. Assoc., 76, 9-15.

SILVERMAN, B. W. (1978). Choosing a window width when estimating a density. Biometrika, 65, 1-12.

TITTERINGTON, D. M. and MILL, G. M. (1981). Kernel-based density estimates from incomplete data. Submitted for publication.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. J. Amer. Statist. Assoc., 69, 169-173.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. J. R. Statist. Soc. B, 38, 290-295.

YANDELL, B. S. (1981). Nonparametric inference for rates and densities for randomly censored data. Ph.D. Thesis, Biostatistics, Univ. of Calif. Berkeley.

DMT/jvs

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2382 — 2. GOVT ACCESSION NO. AD-A116174	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Kernel-Based Density Estimation Using Censored, Truncated or Grouped Data	5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) D. M. Titterington	8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709	12. REPORT DATE May 1982
	13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. *(of this report)* UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

density estimation, kernel method, censoring, truncation, grouping, consistency, imputation

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

Censoring, truncation and grouping represent different but related forms of incompleteness. Methods of producing kernel functions on the incomplete observations are proposed. They involve substituting for or averaging over the incomplete observations. Consistency of the procedures in terms of the criterion of integrated mean squared error is established and optimal choice of smoothing parameter is achieved.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*


DATE
ILME
8